EVALUATION METHODOLOGIES FOR IMAGE RETRIEVAL SYSTEMS

[1]*Ian H. Jermyn,* [2]*Cián W. Shaffrey and* [3]*Nick G. Kingsbury*

[1]Ian.Jermyn@sophia.inria.fr
[1]Project Ariana (CNRS/INRIA/UNSA), INRIA, Sophia Antipolis, France
[2,3]Signal Processing Laboratory, Department of Engineering, University of Cambridge, UK

ABSTRACT

The field of content-based image retrieval is an important one for two reasons. Practically speaking, the oft-cited growth of image archives in many fields, and the rapid expansion of the Web, mean that successful image retrieval systems are fast becoming a necessity. In addition, database retrieval provides a framework within which the important questions of machine vision are brought into focus. This is firstly because successful retrieval is likely to require true image understanding, and secondly because database retrieval provides a potentially objective testbed for image understanding systems. In view of these points, the development of methods for the evaluation of retrieval systems becomes a matter of priority. There is already a substantial literature evaluating various systems, but little high-level discussion of the evaluation methodologies themselves seems to have taken place. This essay proposes a framework within which such issues can be addressed, analyses possible evaluation methodologies, indicating where they are appropriate and where they are not, and critiques some evaluation methodologies used in the literature.

1. INTRODUCTION

It is a commonplace that the growth of the Web and the ever-growing collections of electronic images in many fields renders pressing the need for genuinely content-based image retrieval systems. In addition to this practical necessity however, image retrieval provides a framework within which to view the important problems of machine vision. Successful image retrieval will require genuine image understanding; indeed retrieval is essentially just pullback by the image understanding arrow. As with any field however, progress depends on the ability to evaluate the results of image retrieval (or image understanding) methods in a way that does not depend on the opinion of the evaluator. For collections of images for which there is no well-defined semantics, and hence nothing to which to compare the results of retrieval (a situation that occurs frequently), evaluation methodology becomes a murky area, and little high-level discussion of methodological issues seems to have taken place. This essay tries to shed light on the evaluation of retrieval systems by proposing a framework within which the issues can be described. We perform an analysis of possible evaluation methodologies in scenarios with different degrees of structure, indicating where they are appropriate and where they are not, and critique query by example and techniques associated with it. The analysis is put into practice in [1].

There is already a substantial literature evaluating various retrieval systems. In some ways the closest work in spirit to our analysis is the recent paper by Martin *et al.* [2], although it is not concerned with retrieval as such. They too treat a situation with ill-defined semantics by turning to human subjects, although in a slightly different manner than that advocated here (and used in [1]). Much of the other work in evaluation uses query by example and "relevance" classes of images (see, among many others, [3, 4, 5, 6, 7, 8, 9] and the many references in the reviews [10, 11, 12, 13]). One of the main arguments of this paper is that both these techniques are flawed conceptually, and that used together they give the appearance of objectivity without the substance.

Retrieval systems cover a very broad range of application areas. Some work with very limited 'scenes',[1] and hence with narrowly defined sets of images with precise semantics, while others work with generic scenes whose semantic content seems unbounded. We will bear two examples in mind as we proceed. These examples will serve to make the discussion concrete; they represent two extremes of database usage and evaluation.

L'Institut Géographique National: The first dataset we consider is a collection of aerial images of the Ile-de-France region around Paris created by the Institut Géographique National (IGN), the French Mapping Institute. An example is shown in figure 1. (Each image is 500 by 500 pixels, with a resolution of about 9m/pixel.) There are many classes of statement that one could consider making about such images, but the most important consists of statements about land use. Such statements essentially involve a map from the image domain into a finite set of classes: 'forest', 'urban area', 'agricultural field' and so on. These maps are available: they

This work was supported by EU project MOUMIR (HP108), www.moumir.org

[1]We use the word 'scene' to denote whatever is of interest in the image: this can range from the position and identity of objects in a real (or imagined) 3D scene, to abstract, precisely defined symbols or even camera parameters.

Figure 1: Left: an example of the IGN aerial images © IGN. Right: land use map for part of this image © IAURIF.

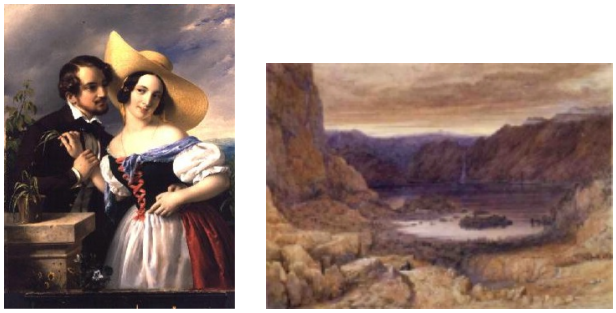

Figure 2: Examples of the BAL images © BAL

have been compiled by the Institut d'Aménagement et d'Urbanisme de la Région d'Ile-de-France (IAURIF). The land use map for part of the image in figure 1 is shown beside it.

Bridgeman Art Library: The second dataset is a collection of scanned images of paintings from the Bridgeman Art Library (BAL). Based in the United Kingdom, BAL is a commercial art library supplying images to magazines, newspapers, designers and others. The images are realistic in intent, but in many cases the colours and forms do not correspond to 'photographic realism'. It is very hard to characterize the queries faced by the staff at BAL. The queries are often phrased at a very high semantic level, and the process of answering queries is complex, involving prolonged interaction with clients. Two example images are shown in figure 2.

2. IMAGE RETRIEVAL SYSTEMS

Database systems are intended for a particular situation. Particular individuals will access the database and try to retrieve images for a particular purpose. The database will either give the individuals what they want with ease and grace, or it will not. This implies that the fullest way to evaluate such systems is to study the performance of the system *in situ*, through the reactions of users, surveys, work rates, and so on, and to come to a conclusion from this data. Whether there is any consistency in the evaluation of different systems across different applications and datasets, or even across different work environments and personnel for the same application and dataset, is an empirical question, to be answered by experiments and not by assumption. If little consistency between the evaluations across different environments were to be found, there would be no well-defined environment-independent sense in which one system could be said to be better than another.

This situation is not very satisfactory. It implies that we can say very little about the performance of image retrieval systems without conducting enormous and often impossible experiments. If we wish to go further in the absence of such experiments, we must *assume* that we can abstract the idea of performance away from the environment in which the software will be used without seriously affecting the evaluations themselves. To make this abstraction, we must find some way to duplicate the evaluation of the 'average relevant user'. By doing so, we are effectively 'integrating out' the variables in which we are not interested, leaving a marginalised performance measure averaged over those variables, in this case the application environment and the diversity of users.

Database Schema

Figure 3 shows a representation of a retrieval system. The figure shows a series of spaces and a number of arrows between them. The arrows can be interpreted either as maps between the spaces themselves, or as maps between the probability measure spaces on those spaces, as we will discuss shortly. We will introduce the components of this diagram one by one.

The 'image space', I, is the space of images. It contains a finite subset: the set of images in the database. (We will not distinguish carefully between I and this subset. This creates no confusion.) Subject to a query, the output of the retrieval system will be a subset of this subset.

The 'semantic space', S, can be thought of as the space of atomic statements we might like to make about the scene for a particular application: statements about a real world scene; the objects represented in a blueprint; the 3D scene represented in a painting; the properties of the painting itself; camera parameters; and so on.

The map h from I to S represents the interpretation of the images by the 'average relevant user'. In general, for each user, their interpretation will take the form of a map between the spaces of probability measures on I and S. In other words, given an image in I, the user's interpretation will map it to a probability distribution over S that represents the uncertainty of the interpretation of that user. To obtain the 'average relevant user', we should then combine the distributions from different users to obtain a distribution over the space of interpretations of many users, and then marginalise over the

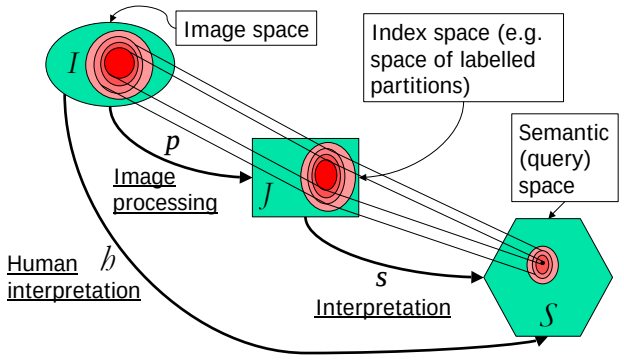

Figure 3: Database schema

different users to obtain a distribution over interpretations. If the entropy of this marginalised conditional distribution is not much higher than the entropy of the distributions for each user individually (there is 'consensus'), then the notion of an 'average relevant user' is useful. Otherwise we are faced with the task of performing large experiments, as discussed at the beginning of this section. If in addition to consensus, the resulting marginalised distribution is narrowly peaked about one value, then the map between probability spaces can usefully be replaced by a map between the spaces themselves. We will talk as though this is the case, although little is changed in the subsequent arguments by relaxing this restriction.

Once I, S and h are defined, we can talk about the ideal output of the retrieval system as follows. A query specifies a point or a set of points in S.[2] Given such a query, the map h can be used to pullback[3] the subset to I, thus specifying a set of retrieved images.[4]

The other space indicated in the figure is the 'index space', J. This, and its associated maps π and σ, constitute a factorization of the machine, as opposed to human, image understanding chain. The map π represents the processing applied to the images in the database to generate indices for retrieval. Usually the high-dimensional image space is projected to a much lower-dimensional space that it is hoped nevertheless still captures the relevant information about the images. The latter phrase means that there exists a map τ such that $h = \tau\pi$. This cannot happen in particular, if two images map to the same point in J under π that are mapped to different points in S under h. In other words, the index space and the image processing map must be 'fine enough'. The semantics map, σ, represents further processing that maps the index space to statements about the image. The ideal situation is that $\sigma\pi = h$. If this were the case, then machine retrieval would follow the same lines as ideal human retrieval: pullback from a query using the arrow $(\sigma\pi)^* = \pi^*\sigma^*$. One may wonder why we need to factorize the map $\sigma\pi$ and create the space J. The answer is that construction of a complete map $\sigma\pi$ that approximates h is often impossible at the present state of development. The space J thus represents as far as we can presently go along the line between images and semantics; J does not equal S except in a limited number of cases.

Knowledge Scenarios

Problems arise in building and evaluating retrieval systems because one or the other of the above quantities is difficult, if not impossible, to characterize algorithmically. These problems have a well-defined order, in the sense that the inability to solve a problem earlier in the list renders the subsequent problems moot. Each stage in this list will be called a 'knowledge scenario' or 'KS'.

KS 1: In this KS, we cannot characterize S explicitly. Consequently, neither h nor σ are characterizable explicitly. It is usually only possible to construct S if a limited number of anti-atoms can be found in terms of which to express the statements as conjunctions. In the case of the IGN images, for example, conjunctions of statements of the form: "such-and-such region in the image domain corresponds to land use of such-and-such type" are enough to express all relevant queries. No such simple characterization exists for the images from the BAL dataset, and indeed the semantic space seems unbounded. The BAL dataset and the queries typically made of it are a good example of this first KS.

KS 2: In this KS, we know how to characterize S, but we do not know how to characterize h. Difficulties with characterizing h fall into two classes. There may be consensus among users about the semantics in a probabilistic sense, while still being a great deal of uncertainty about the interpretations. In this case, h must be described using conditional probabilities. It may be difficult to obtain the information necessary to describe this distribution. In addition, I and S may be too large to allow explicit construction of the map, requiring prohibitive resources of money or time. In the IGN case, exhaustive enumeration is possible. There is one image giving the value of h for every image in the database; these were created by human operatives. It is however easy to imagine increasing the complexity of the semantics or the number of

[2] In practice, this subset may be specified in several ways; we suppose that the query is specified directly. Consequently we do not examine such techniques as 'relevance feedback', which can be viewed as methods for perfecting the query. This then excludes the discussion of such performance characteristics as how fast a user may reach their final query, and of what limits the feedback process may put on the trajectories possible in the semantic space.

[3] For an arrow $f : A \to B$, we use the notation $f^* : 2^B \to 2^A$ to indicate pullback: $\forall Y \subset B : f^*(Y) = \{a \in A : f(a) \in Y\}$, and the notation $f_* : 2^A \to 2^B$ to indicate push forward: $\forall X \subset A : f_*(X) = \{f(a) \in B : a \in X\}$.

[4] Many methods introduce a metric on S. We consider the use of a metric either as specifying a less restrictive query or as a probabilistic version of the current discussion. Nothing is altered.

images in such a way that exhaustive enumeration would become impossible. We might then know the statements that we wish to make, but be unable to ascertain whether or not they are true of a given image.

KS 3: In this KS, we know how to characterize S and h, but we cannot construct a version of the arrow $\sigma\pi$ that approximates h. This KS is a little different from the previous two. There, we could not characterize what we wanted to retrieve. Here, we can characterize it, but we cannot hope to duplicate it. Our attempts at retrieval are then doomed to failure. An example is the following. Assume we have a set of images and that we wish to make statements of the form "In the scene that generated this image, a human being occupied a volume that projected to region R in the image domain". The semantics is very clear, and we can construct h simply by inspection, or through prior knowledge of the scenes from which the images were generated. It is however a difficult task to do this automatically for general classes of images.

KS 4: In this KS, we can characterize both S and h, and we can construct reasonably successful maps $\sigma\pi$. The IGN dataset, coupled with queries about land use, form a good example of this KS. We have a well-defined semantic space, mentioned in KS1, a well-defined arrow h, given by the IAU-RIF images, and segmentation algorithms that do a reasonable job of partitioning the image domains into the correct subsets.

Query by Example: It is unfortunately all too often the case that we find ourselves in KS1. One common response to this is to attempt to circumvent the need for a semantics as follows. In the absence of any well-defined maps, one instead assumes that the user interprets images via an unknown map h to an unknown S, and that he has in mind an unknown query $q \subset S$. One then allows the user to select a small subset $i \subset I$ of images from the database that 'represent' the unknown query q, which means in principle that $i \subset h^*(q)$. Now that one has a set of images to work with rather than a query, one can apply the arrow $\pi^*\pi_*$ to generate a set of retrieved images in I. This is known as 'query by example' ('QBE'). QBE removes the need for the unknown quantities S and h, and hence σ, by employing the user himself to translate from the unknown query to a set of 'example' images.

Note that if there exists an arrow σ such that $h = \sigma\pi$, then

$$h = \sigma\pi \Rightarrow h^* = \pi^*\sigma^* \Rightarrow \sigma^* = \pi_* h^* \Rightarrow h^* = \pi^*\pi_* h^*. \tag{1}$$

It then follows that $i \subset \pi^*\pi_*(i) \subset h^*(q)$, so that if σ exists, the retrieval process will return more images with the same semantics. In fact, the existence of a σ such that $h = \sigma\pi$ means that π divides I into equivalence classes that are a refinement of those generated by the unknown h: images i and i' such that $\pi(i) = \pi(i')$ necessarily satisfy $h(i) = h(i')$. QBE raises a number of difficulties. Note that a similar procedure does not work as soon as we move to KS2. Once we know the semantic space, the user's undisclosed interpretation of the images is open to question. A second problem concerns the unknown query. The subset of images selected, i, will be a subset of $h^*(q)$ for a great many queries q and arrows h. How does the retrieval system know which of these is intended by the user? Clearly it cannot. What then do the images it retrieves represent? A third difficulty concerns the existence of the arrow σ. In most cases, it is obvious that such an arrow does *not* exist, in which case equation 1 will be incorrect: the method cannot produce the correct results. What then are we to make of the claims of successful retrieval reported in the literature? We leave further consideration of these issues until we discuss the evaluation of QBE in section 3.1.

3. EVALUATION CONTEXTS

We move from the structure of a retrieval system and the difficulties involved in its construction, to consider its evaluation. Evaluation always takes the form of a comparison between two arrows with common domain D and co-domain C: a 'reference arrow', which describes the ideal behaviour, and a 'test arrow'. We will call the co-domain C the 'evaluation context' or 'EC'. We introduce a probability measure μ on D and a 'utility/loss function' ρ on C (or 2^C). Both these quantities are to be decided by the evaluator. In our case, the measure on D might correspond to the frequency with which certain queries are put to the system. The comparison between two arrows a and b then takes the form

$$\Upsilon(a, b) = \mu(\rho(a \times b)\Delta) \tag{2}$$

where: a is the reference arrow, which will always involve h; b is the test arrow, which should not involve h; Δ is the diagonal map $D \to D \times D : d \mapsto (d, d)$; and μ integrates its argument. The value of Υ is thus a measure of how well/badly we are doing by using b instead of a. A score of 0 would indicate perfection: the arrows a and b are the same for the purposes in which we are interested. (This may not imply $a = b$.)

To specify an evaluation method it is necessary to specify C, D, a and b, but in the case of figure 3, a and b are specified once C and D are given, and in fact there is always an obvious choice of D also. We can therefore concentrate on the EC's. There are three of them: the 'image context' (which we will also call the 'retrieval context'), the 'indexing context' and the 'semantic context'. Whether a given EC can be used depends on which of the arrows in figure 3 are available, since computation of $\Upsilon(a, b)$ is clearly impossible if we cannot compute a and b. We will thus see that different EC's are appropriate in different KS's. We now define the three EC's in more detail.

Semantic Context $C = S$: The natural choice for D is I. The arrows being compared are h and $\sigma\pi$. Equation 2 then becomes:

$$\Upsilon = \sum_{i \in I} \rho_S(h(i), \sigma\pi(i))\mu_I(i) \tag{3}$$

Retrieval Context $C = I$: This is the EC most often evoked in the literature, presumably because it seems to offer the most direct connection to the 'average relevant user' and to retrieval performance. It is rarely used in the complete form presented here; more often, a version appropriate to the assumptions of QBE is used. We will discuss this further in section 3.1. The natural choice of D is S, and the arrows being compared are h^* and $\pi^*\sigma^*$. These map points of S to subsets of I, so that we must define a utility/loss function on 2^I. Using a typical choice of utility/loss, equation 2 becomes

$$\Upsilon = \sum_{q \in S} \frac{|h^*(q) \cap \sigma\pi^*(q)|}{|h^*(q)|}\mu_S(q) \tag{4}$$

This utility/loss function is known as "recall". It normalises the number of 'successful' images found by the number in the database. Based on another obvious normalisation, one can also define the "precision", given by

$$\Upsilon = \sum_{q \in S} \frac{|h^*(q) \cap \sigma\pi^*(q)|}{|\sigma\pi^*(q)|}\mu_S(q) \tag{5}$$

Indexing Context $C = J$: The natural choice of D is I, while the arrows are π and $\sigma^* h$. The form of the score in this case is

$$\Upsilon = \int_{i \in I} \rho_J(\sigma^* h(i), \pi(i))\mu_I(i) \tag{6}$$

The domain of this Υ is $2^J \times J$. Note that we do not need separate characterizations of h and σ, since only the combination $\sigma^* h$ appears in equation 6. We must however make an assumption. In order to compare the performance of different methods, we need a common J. This limits the range of applicability of this EC to the comparison of systems that share, at least to some degree, the same J. Often we can generate a common J by removing structure. Suppose for example that the index spaces of various systems consist of partitions of the image domain labelled by the values of different features. A direct comparison is impossible, but by keeping only the common structure (the image domain partitions), a comparison is enabled. The effect of this is to 'coarsen' the semantics we can hope to capture, since it will group each index space into equivalence classes.

3.1. Use of Different EC's

Having defined the EC's, we can now look more closely at when they can be applied, and in particular, in which KS's they are relevant. It will turn out that lack of explicit knowledge of certain maps need not hinder us if we can define them implicitly. We look at the KS's one by one, and at QBE separately, since it has properties peculiar to the assumptions used in its definition.

KS 4: In this case, the natural EC to use is the semantic context. All the arrows are defined, and we need only define a utility/loss function on S. Since by assumption we can construct arrows $\sigma\pi$ that come close to h, we can expect reasonably high scores in our evaluations. In addition, the other EC's are guaranteed to give good results if we have an image processing method $\sigma\pi$ that duplicates h. Thus the retrieval context, although apparently well-adapted to this situation, is not really needed, and indeed is harder to use: it is more difficult to define meaningful utility/loss functions on I than on S.

KS 3: In this case, we cannot use the semantic context, or rather use of it is meaningless. The arrows $\sigma\pi$ that we know how to build come nowhere near duplicating h, so that claiming victory for one method over another is really beside the point. The same consideration applies to the retrieval context. In the case of the indexing context, one might hope that we could define an arrow $\sigma^* h$ without explicitly defining σ, but the fact that we already know h does not allow us the freedom to do so. Thus KS3 turns out to be impossible to evaluate. This highlights the somewhat dubious nature of evaluation in the KS's to come, in which we have even less knowledge. The lack of constraint allows us to make progress by making simplifying assumptions, but KS3 makes it clear that we should be wary of drawing hasty conclusions from any apparent success such evaluations may produce.

KS 2: We cannot use the semantic and retrieval contexts since these rely on knowledge of h, which we lack. Similarly to KS3, we cannot make assumptions about $\sigma^* h$, since although we do not know h, we do know S. Again evaluation is not possible.

KS 1: As in the previous KS, the semantic and retrieval contexts are eliminated. The indexing context however is not eliminated so easily. In the indexing context we do not need the semantic space explicitly, nor do we need separate characterizations of h and σ; we need only the combination $\sigma^* h$. Since there are now no constraints on the individual arrows making up the combination, we are freer to try to characterize this arrow in another way. The natural way to do this is through the use of human subjects. This is a subtle point: we are saying that although it is not possible to characterize the semantics of images directly, nevertheless we can gain access to some knowledge about those semantics by looking instead at the results in J that might generate those semantics correctly.

How should we set about using human subjects to characterize this arrow? Clearly human image understanding does not generate points in J. We cannot therefore ask human sub-

jects to tell us their interpretations and use these as a characterization of h, as we might do if we had a well-defined semantic space. We can however ask human subjects to evaluate directly the points in J that are generated by the arrow π, thus characterizing h implicitly. Absolute scoring of individual arrows will not do, since the meaning of the absolute scores will be very unclear, but a comparison of the outputs of the different π arrows stemming from different methods is nevertheless possible. The subject can be asked which of the two representations generated by two different methods is more meaningful. The most interesting result of such an evaluation is the existence or not of consensus among subjects. Its existence indicates that there may be an underlying 'fundamental' image semantics to which we can gain access via such experiments. We discuss this methodology further when we discuss the BAL dataset in section 4.

QBE: We note first that there is only one arrow so far in QBE, $\pi^*\pi_*$, and that consequently there is nothing to evaluate. In order to proceed further, a second, reference arrow is needed. To this end, an arrow $\hat{h}$ is introduced, that notionally maps I to some semantics $\hat{S}$. (Neither $\hat{h}$ nor $\hat{S}$ are *a priori* the same as the quantities h and σ that are supposed to generate the 'example' images in the retrieval process itself.) The arrow $\hat{h}$ is not described directly, since to do so would require a definition of $\hat{S}$. Only the combination $\hat{h}^*\hat{h}_*$ is described, by giving the partition of I into equivalence classes sharing equal values of $\hat{h}$. The equivalence class of images to which a given image belongs is known as the set of images "relevant" to the given image.

The introduction of $\hat{h}$ enables the comparison of the arrows $\hat{h}^*\hat{h}_*$ and $\pi^*\pi_*$, the arrow defining retrieval in QBE, since both map 2^I to itself. One can use appropriately adapted versions of equations 4 and 5 to compare the number of "relevant" images retrieved to the number of "relevant" images in the database or to the number of retrieved images. The way this is done in practice is the following. The database of images is divided into groups, supposed to represent the arrow $\hat{h}^*\hat{h}_*$. These groups are typically based on the 'generic name' of the 'most prominent object' in the image. To test the retrieval abilities of the system, an image or set of images i is pulled back by the arrow $\pi^*\pi_*$, giving a set of retrieved images, which are then compared to the set of images "relevant" to i. The results of these tests are sometimes quite remarkable. Recall and precision values above 90% are not at all unusual. Are we really this good at content-based image retrieval?

What does it mean that the images retrieved and the relevance classes into which I is divided agree so closely? In its raw form, it means that the arrow π has managed to duplicate the grouping of I into equivalence classes. In itself, this is not that impressive of course. Given enough parameters, any classification can be duplicated. The inference from the results, given the grouping of I, is however closer to the following: "An image of a horse was used as a query, and the retrieved images consisted of almost all the horses in the database and very little else. Thus the method captures the image semantics.". Let us analyse this statement.

The first point to note is that h and $\hat{h}$ are not necessarily the same. If they are not the same, then we would not expect retrieval based on $\pi^*\pi_*$ to agree with retrieval based on $\hat{h}^*\hat{h}_*$. Thus, while we may choose to imagine that the user was actually looking for horses, he may have been looking for images with any of a number of other interpretations. The user can change his mind about his interpretation at will, while still using the same set of 'example' images. Thus the fact that the retrieved set of images consists of horses may or may not be a success, depending on whether h and $\hat{h}$ are equal. Calling "obvious" the interpretation that renders the retrieval a success, does not solve this problem.

The second point concerns the existence of an arrow σ such that $h = \sigma\pi$. (We now assume that $\hat{h} = h$.) If such an arrow exists then, as shown above, we will have that $h^* = \pi^*\pi_*h^*$. This means that precision will be 100%. If further, we have that σ is a bijection, then recall will be 100% also. The values of recall and precision reported in the literature suggest that this situation is close to being reached. This means that π divides the space of images into the same equivalence classes as h. Since h is never specified, it is of course unclear what this actually means. The clear implication however is that π is actually classifying images into the 'generic classes' of the 'most prominent object' in the image, that is, horses, cars,.... This is remarkable in methods that contain no models of these objects, and which sometimes use the crudest of global features. In fact, it is apparent that π is achieving no such thing. What then to make of the success of the retrieval experiments? Clearly, I must possess a remarkable structure. Firstly, the 'generic name' of the 'most prominent object' is closely correlated with the low-level features typically involved in π, and secondly, the images are well-clustered in J. Since neither is true in general, we are forced to conclude that the I being used to test the methods is very special, and that little can be made of the results reported.

4. SPECIFIC APPLICATION DOMAINS

We turn now to a consideration of the application domains that we described in section 1: the IGN and BAL datasets, and how they fit into the above analysis. The evaluations discussed in this section are in progress. The results of the BAL study are reported in [1]. The results of the IGN study will be published at a later date.

IGN Datasest: For the IGN dataset, the semantic space S is given by the conjunction of statements such as those mentioned in KS1. In addition, the actual land use is known, having been compiled from existing maps and field studies. Thus h is characterisable in the form of a ground truth land

use image for each image in the database. In addition, the semantic space is such that we can construct reasonably accurate versions of $\sigma\pi$. The nature of the images in the IGN dataset creates this possibility: at the resolution of the images, the scene is more or less a flat two-dimensional surface, with texture 'painted' onto it, and in addition the different types of land use seem characterisable in terms of texture descriptors and other low-level image features. We are thus in KS4.

The way forward now depends on exactly what we want to test. If we want to test the performance of the image processing methods with respect to retrieval, then, according to the above analysis, the ideal EC is the semantic context. Each method will produce a partition of the image domain labelled with various land uses, and we can then compare, based on a metric such as the percentage of land area misclassified, the performance of the methods.

Another possibility is that we wish to test segmentation methods in a way that is independent of classification. In this case, we are interested only in the partitions and not in the labels attached to them, so that we can simplify S by 'forgetting' the labels, and compare partitions directly. We are still in the semantic context, but with a reduced semantics. This would enable the evaluation of unsupervised as well as supervised segmentation methods.

BAL Dataset: For the BAL dataset, the semantic space seems unbounded. The Bridgeman Art Library has to deal with queries of a very abstract nature, whose relation to image properties is extremely complicated, involving a great deal of cultural knowledge. In addition, individuals may not be clear about their own interpretation, and it is almost certain that there will not be consensus over some of the statements one might like to make about the images.

We can however simplify this situation somewhat. Whatever the nature of the statements we wish to make about the images, it seems clear that they will require as a necessary, although by no means sufficient input, the identification of the 'principle objects' in the image. We can therefore reduce our semantics somewhat by restricting ourselves to disjunctions and conjunctions of statements of a form rather similar to those used for the IGN images: "Such-and-such region of the image contains such-and-such (named) object". Unfortunately, this is still too broad. Statements about the BAL images contain a far larger set of objects than the IGN images, so many in fact that it is not feasible to list them all. We could give a fixed list of objects and define a semantics in those terms, but this is too restrictive. The absence of any well-defined semantic space means we are in KS1. The semantic and retrieval contexts are thus ruled out.

At this point, we could try to invoke the assumptions of QBE, and at the same time classify the images in the BAL dataset into "relevance" classes as described above. The drastic nature of the QBE assumptions is exposed once we think about applying them to image and semantic spaces as complex as those of the BAL dataset. Most images from the BAL dataset do not clearly specify any query, and any attempt to categorize the images into "relevance" classes for the purpose of evaluation seems completely arbitrary.

Instead, we turn to the indexing context, which, free as it is from the need to characterize S and h, has not yet been ruled out. We assume a working hypothesis: that there does exist human consensus about what might be called 'fundamental' image segmentations. (As recent work has shown [2], consensus may well exist at least for limited classes of images.) We then proceed as follows. In order to compare a number of segmentation procedures, they must share a common J. This we ensure by defining J as a space of (unlabelled) partitions of the image domain. The above assumption amounts to assuming that for each image, there is human consensus about a semantic interpretation that includes as part of its definition an image domain partition. We can thus ask individuals to 'score' the output of various segmentation algorithms by comparison with the original image and, in practice, to avoid the arbitrariness involved in an absolute scoring system, by comparison with each other.

Note that what is being tested is not simply the performance of different methods, but the very existence of a consensus about the interpretation of the images involved. The existence of such a consensus is far from obvious, and is arguably a more interesting question than the results of the evaluations themselves.

5. CONCLUSION

The ease of application of QBE, coupled with the notion that there is something 'special' about image data, seems to have created the impression that for image retrieval it may in principle be preferable to query by text. It is undeniable that there is something special about image data, at least when compared to text or speech retrieval. The segmentation of sound into words leaves one with text, which is composed of atoms that exist already at the semantic level: one does not need to 'name' words. The number of atoms is limited and they are known *a priori*. In image understanding, the number of semantic atoms is vastly greater, and the correspondence of segmented regions (for example) with semantics is not clear. Coupled with higher dimensionality, which allows geometry to intrude, and the projective nature of image formation, we see that image understanding is vastly harder than speech processing. We reject the notion however that these differences require a qualitatively different approach to image retrieval. This is because semantics seems to us inherently linguistic. This is supported by the psychophysical experiments performed using the *PicHunter* system [14]. QBE in which the meaning of the example is not clarified both to the user and to the system by linguistic cues that specify a

query in a well-defined semantic space, results in confusion, as the above analysis shows. We believe that it cannot be a substitute for the characterization of a semantic space, an interpretation arrow h, and a genuine image processing arrow $\sigma\pi$ from I to S. Our current inability to construct such arrows in many cases of interest should not be disguised by lack of methodological clarity.

The evaluation method that combines "relevance" classes with QBE suffers from a number of serious drawbacks, not the least of which is the appearance of objective evaluation without the substance. 'Success' in such evaluations is less an expression of the ability of the image processing system involved, and more a statement about the distribution of the images in the database. Image retrieval systems *can* be evaluated within the semantic and retrieval contexts, but only if we have a characterization of S and h. In the absence of such characterizations, we are forced to move to the indexing context, and to perform psychological experiments with human subjects in order to evaluate systems.

Many of the above considerations apply directly to the evaluation of image processing methods in general. The reason for this is clear: image retrieval is in essence pullback by the image interpretation arrow $\sigma\pi$. The accuracy of the retrieval is entirely dependent on the accuracy of this interpretation. In cases where no well-defined semantics is available, the only available evaluation method for image processing systems in general is the psychovisual one proposed above for the BAL dataset. The process involved in such experiments is similar to that used in the 'eye of the beholder' method of evaluation that is all too common in image processing. The difference is that properly designed experiments take into account a large number of different images and a range of different users in order to test the idea of a consensus and produce an evaluation if such a consensus exists.

Acknowledgements: The authors would like to thank the Institut Géographique National, the Bridgeman Art Library and the Institut d'Aménagement et d'Urbanisme de la Région d'Ile-de-France for the use of the images in this document.

6. REFERENCES

[1] C. W. Shaffrey, I. H. Jermyn, and N. G. Kingsbury, "Psychophysical evaluation of image segmentation algorithms," Submitted to ACIVS 2002.

[2] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8^{th} IEEE Int'l Conf. Comp. Vis.*, Vancouver, Canada, July 2001.

[3] C. Meilhac and C. Nastar, "Relevance feedback and category search in image databases," in *ICMCS*, 1999, vol. 1, pp. 512–517.

[4] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A power tool for interactive content-based image retrieval," *IEEE Trans. Circuits and Video Tech.*, vol. 8, no. 5, pp. 644–655, 1998.

[5] R. Brunelli and O. Mich, "Image retrieval by examples," *IEEE Trans. Multimedia*, vol. 2, no. 3, pp. 164–171, 2000.

[6] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, "Blobworld: A system for region-based image indexing and retrieval," in 3^{rd} *Int'l Conf. Visual Information Systems*. 1999, Springer.

[7] S. Ardizzoni, I. Bartolini, and M. Patella, "Windsurf: Region-based image retrieval using wavelets," in *DEXA Workshop*, 1999, pp. 167–173.

[8] D. McG. Squire, H. Müller, W. Müller, S. Marchand-Maillet, and T. Pun, *Design and Evaluation of a Content-based Image Retrieval System*, chapter 7, pp. 125–151, Idea Group Publishing, 2001.

[9] E. Müller, W. Müller, S. Marchand-Maillet, D. McG. Squire, and T. Pun, "A web-based evaluation system for content-based image retrieval," in *Proceedings of the ACM Multimedia Workshop on Multimedia Information Retrieval (ACM MIR 2001)*, Ottawa, Canada, 2001, pp. 50–54.

[10] K. Bowyer and P. Flynn, "Content-based image retrieval at the end of the early years," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[11] J. R. Smith, "Image retrieval evaluation," in *IEEE Workshop on Content-Based Access of Image and Video Libraries*, June 1998.

[12] Y. Rui, T. Huang, and S-F. Chang, "Image retrieval: Current techniques, promising directions and open issues.," *Journal of Visual Communication and Image Representation*, March 1999.

[13] H. Müller, W. Müller, D. McG. Squire, S. Marchand-Maillet, and T. Pun, "Performance evaluation in content-based image retrieval: Overview and proposals," *Pattern Recognition Letters (Special Issue on Image and Video Indexing)*, vol. 22, no. 5, pp. 593–601, 2001.

[14] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, "The Bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments," *IEEE Trans. Image Proc.*, vol. 9, no. 1, pp. 20–37, 2000.